Surveying the NIH Biomedical Informatics and Computational Biology (BICB) Portfolio through Supervised Machine Learning

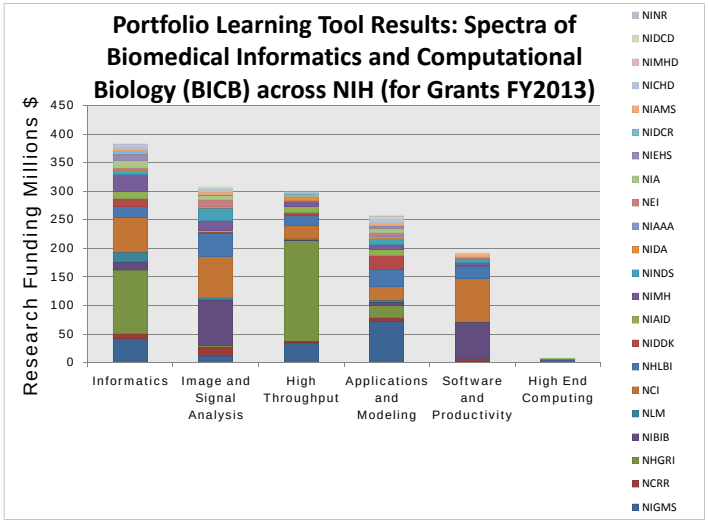
Peter Lyster[1], Calvin Johnson[2], William Lau[2], Kelley Smith[1]

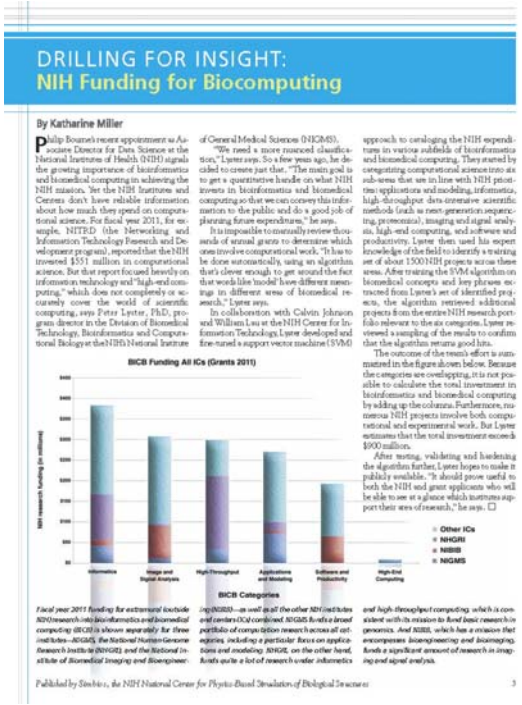
1 – Division of Biomedical Technology, Bioinformatics, and Computational Biology, NIGMS

2 – Division of Computational Bioscience, CIT

ABSTRACT

We describe an approach for classifying NIH research funding dealing with "Biomedical Informatics and Computational Biology" (BICB). In doing so, we describe a method to obtain an inventory, including dollar amounts, of grants and contracts. The approach we have adopted involves first developing a set of parsimonious categories that describe the types of BICB research projects that are funded by NIH: applications and modeling, informatics, image and signal analysis, high throughput tools, software and productivity, biostatistics, and high end computing. One of us (PML) then acted as a 'rater' who identified a gold-standard set of projects out of the broad NIH portfolio of research grants and assessed to be a best fit to the BICB classes; this is called a 'training set'. In the course of our research we developed a support-vector machine based classifier, here referred to as the Portfolio Learning Tool (PLT), which was developed for retrieving a complete inventory of grants and contracts based on the training set. The PLT is a flexible and extensible framework for retrieving specific research inventories. In the present context it is used only to develop a BICB inventory, however the methodology has broad applicability. In the inventory of the BICB area, our findings show that NIGMS funds a significant portfolio of the 'Applications and Modeling' category in Biomedical Informatics and Computational Biology (BICB), followed by the 'Informatics' and "High-Throughput' categories. NHGRI provides considerable funding for research in the 'High-Throughput' and 'Informatics' BICB categories, as well as supporting a smaller portion of the 'Applications and Modeling' category. NIBIB provides substantial support in the BICB areas of 'Image and Signal Analysis' and 'Software and Productivity' categories. This knowledge allows for greater understanding of the primary focus and missions of IC portfolios.




DRILLING FOR INSIGHT:
NIH Funding for Biocomputing

By Katharine Miller

Philip Bourne's recent appointment as Associate Director for Data Science at the National Institutes of Health (NIH) signals the growing importance of bioinformatics and biomedical computing in achieving the NIH mission. Yet the NIH Institutes and Centers don't have reliable information about how much they spend on computational science. For fiscal year 2011, for example, NITRD (the Networking and Information Technology Research and Development program) reported that the NIH invested $551 million in computational science. But that report focused heavily on information technology and "high-end computing," which does not completely or accurately cover the world of scientific computing, says Peter Lyster, PhD, program director in the Division of Biomedical Technology, Bioinformatics and Computational Biology at the NIH's National Institute of General Medical Sciences (NIGMS).

"We need a more nuanced classification," Lyster says. So a few years ago, he decided to create just that. "The main goal is to get a quantitative handle on what NIH invests in bioinformatics and biomedical computing so that we can convey this information to the public and do a good job of planning future expenditures," he says.

It is impossible to manually review thousands of annual grants to determine which ones involve computational work. "It has to be done automatically, using an algorithm that's clever enough to get around the fact that words like 'model' have different meanings in different areas of biomedical research," Lyster says.

In collaboration with Calvin Johnson and William Lau at the NIH Center for Information Technology, Lyster developed and fine-tuned a support vector machine (SVM) approach to cataloging the NIH expenditures in various subfields of bioinformatics and biomedical computing. They started by categorizing computational science into six sub-areas that are in line with NIH priorities: applications and modeling, informatics, high-throughput data-intensive scientific methods (such as next-generation sequencing, proteomics), imaging and signal analysis, high-end computing, and software and productivity. Lyster then used his expert knowledge of the field to identify a training set of about 1500 NIH projects across these areas. After training the SVM algorithm on biomedical concepts and key phrases extracted from Lyster's set of identified projects, the algorithm retrieved additional projects from the entire NIH research portfolio relevant to the six categories. Lyster reviewed a sampling of the results to confirm that the algorithm returns good hits.

The outcome of the team's effort is summarized in the figure shown below. Because the categories are overlapping, it is not possible to calculate the total investment in bioinformatics and biomedical computing by adding up the columns. Furthermore, numerous NIH projects involve both computational and experimental work. But Lyster estimates that the total investment exceeds $900 million.

After testing, validating and hardening the algorithm further, Lyster hopes to make it publicly available. "It should prove useful to both the NIH and grant applicants who will be able to see at a glance which institutes support their area of research," he says. □

BICB Funding All ICs (Grants 2011)

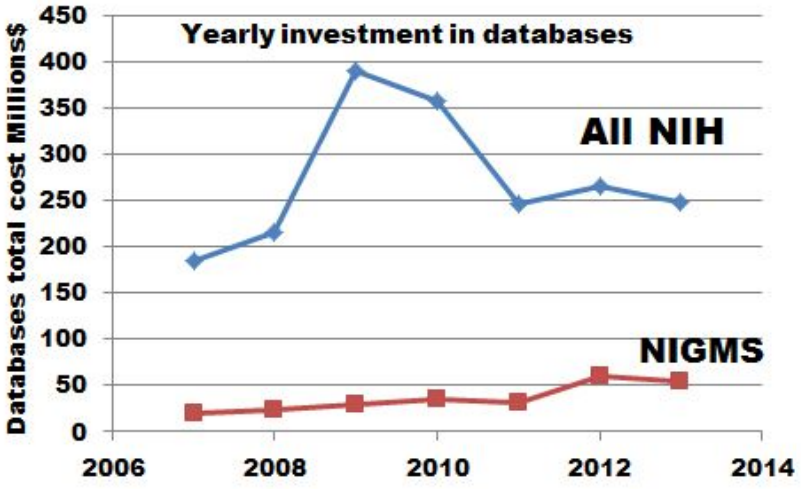
BICB Categories

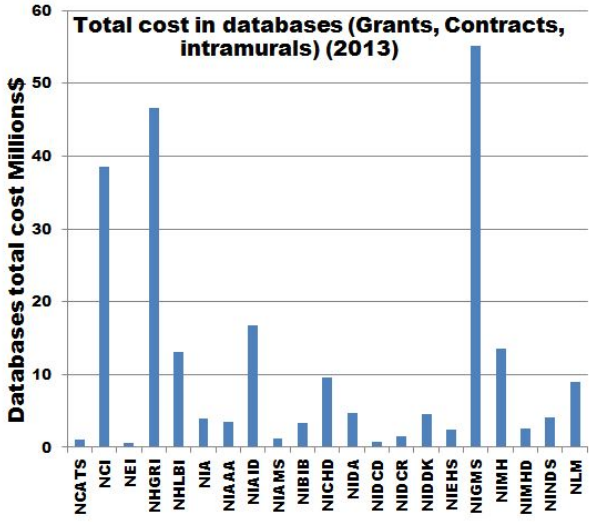
Other ICs
NHGRI
NIBIB
NIGMS

Fiscal year 2011 funding for extramural (outside NIH) research into bioinformatics and biomedical computing (BICB) is shown separately for three institutes—NCBI, the National Human Genome Research Institute (NHGRI), and the National Institute of Biomedical Imaging and Bioengineering (NIBIB)—as well as all the other NIH institutes and centers (ICs) combined. NCBI's broad portfolio of computation research across all categories, including a particular focus on applications and modeling, NHGRI, on the other hand, funds quite a lot of research under informatics and high-throughput computing, which is consistent with its mission to fund basic research in genomics. And NIBIB, which has a mission that encompasses bioengineering and bioimaging, funds a significant amount of research in imaging and signal analysis.

Published by Sunbio, the NIH National Center for Physics-Based Simulation of Biological Structures

Yearly Investment in Databases at GM vs. NIH





FINDINGS

In addition, we developed a training set to identify all databases from among research, contract, and intramural projects at NIH. The results are presented in the figures on the left and show that NIGMS, NHGRI, and NCI have the highest investment in databases for FY2013.

The increase in the yearly investment in databases from 2008-2009 is believed to be due to the ARRA investment, and is something that we plan on looking into moving forward.

We continue to work on validation across BICB and database retrievals, as it is a living process.